ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
マルチモーダル情報を活用するタンパク質言語モデル

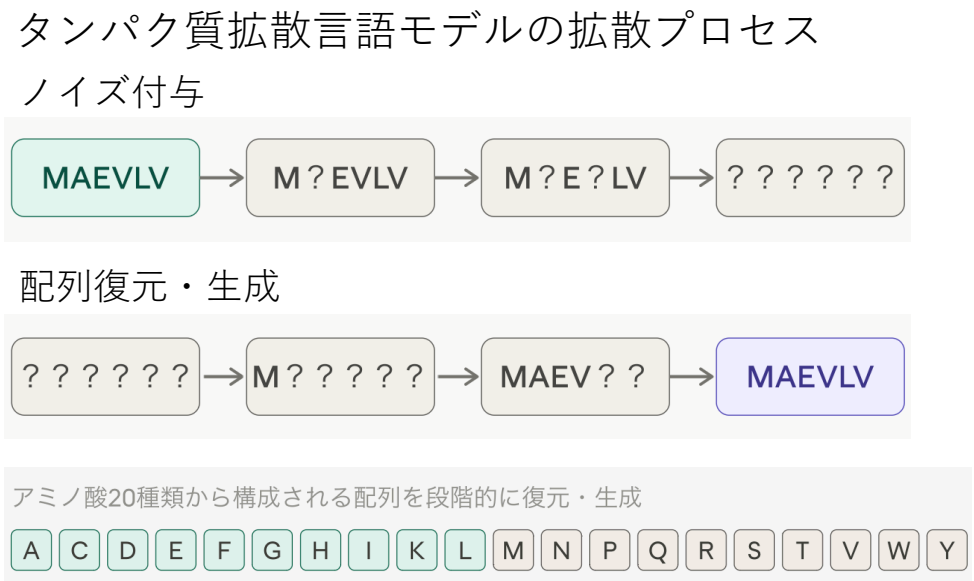
実施時期：2025年4月~2026年3月
所属機関名：産業技術総合研究所
代表者氏名：浅田 真生

成果概要：

本課題では、タンパク質配列データを基盤として、新規配列を生成する拡散モデルの研究開発基盤を構築した。具体的には、大規模データベース（約4,200万配列）を用いた拡散型タンパク質言語モデルのベースライン実装および再現実験を行い、計算基盤上での安定した学習・評価パイプラインを整備した。

成果のポイント：

① 拡散型タンパク質言語モデルの実装・再現
タンパク質配列はアミノ酸配列として表現でき、テキスト生成と類似の枠組みで扱える。本研究では、大規模配列データで学習された拡散型タンパク質言語モデルを計算基盤上に実装し、再現実験を行った。全アミノ酸をマスクした状態から段階的に復元する生成過程を再現し、安定した学習・推論パイプラインを整備した。これにより、今後の研究に向けた基盤を構築した。
② タンパク質配列生成の評価指標の整備
生成配列の品質を評価するため、複数の観点からなる評価基盤を整備した。具体的には、アミノ酸頻度分布の一致度、配列多様性、新規性、ユニーク率の4指標を用いた評価フレームワークを構築した。さらに、外部モデルによる疑似パープレキシティを導入し、生成配列の自然性を定量的に評価可能とした。
③ テキスト生成手法のタンパク質領域への適用可能性の分析
テキスト生成で用いられる手法の適用可能性を検討し、タンパク質配列特有の性質を分析した。その結果、語彙サイズが小さいため特定のアミノ酸への偏りが生じやすいことを確認した。これを踏まえ、パラメータ調整による短期的対応、生成多様性を考慮した手法拡張、物理化学的特性を取り入れた評価設計など、段階的な改良方針を整理した。



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：